Zhai *et al. BMC Medical Research Methodology* (2017) 17:12
DOI 10.1186/s12874-017-0291-y

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Ridle for sparse regression with mandatory covariates with application to the genetic assessment of histologic grades of breast cancer

Jing Zhai, Chiu-Hsieh Hsu and Z. John Daye*

Abstract

Background: Many questions in statistical genomics can be formulated in terms of variable selection of candidate biological factors for modeling a trait or quantity of interest. Often, in these applications, additional covariates describing clinical, demographical or experimental effects must be included a priori as mandatory covariates while allowing the selection of a large number of candidate or optional variables. As genomic studies routinely require mandatory covariates, it is of interest to propose principled methods of variable selection that can incorporate mandatory covariates.

Methods: In this article, we propose the ridge-lasso hybrid estimator (ridle), a new penalized regression method that simultaneously estimates coefficients of mandatory covariates while allowing selection for others. The ridle provides a principled approach to mitigate effects of multicollinearity among the mandatory covariates and possible dependency between mandatory and optional variables. We provide detailed empirical and theoretical studies to evaluate our method. In addition, we develop an efficient algorithm for the ridle. Software, based on efficient Fortran code with R-language wrappers, is publicly and freely available at https://sites.google.com/site/zhongyindaye/software.

Results: The ridle is useful when mandatory predictors are known to be significant due to prior knowledge or must be kept for additional analysis. Both theoretical and comprehensive simulation studies have shown that the ridle to be advantageous when mandatory covariates are correlated with the irrelevant optional predictors or are highly correlated among themselves. A microarray gene expression analysis of the histologic grades of breast cancer has identified 24 genes, in which 2 genes are selected only by the ridle among current methods and found to be associated with tumor grade.

Conclusions: In this article, we proposed the ridle as a principled sparse regression method for the selection of optional variables while incorporating mandatory ones. Results suggest that the ridle is advantageous when mandatory covariates are correlated with the irrelevant optional predictors or are highly correlated among themselves.

Keywords: Gene expression analysis, Lasso, Linear models, Penalized regression, Ridge, Variable selection

*Correspondence: zhongyindaye@gmail.com
Epidemiology and Biostatistics Department, University of Arizona, Tucson, USA

Background

Many essential problems in statistical genomics may be formulated in terms of variable selection of candidate biological factors for modeling of some trait or quantity of interest [1–3]. Often, additional covariates describing clinical, demographical, or other experimental factors must be included a priori as mandatory covariates while allowing the selection of possibly a large number of candidate or optional variables. Substantial progress has been made recently in the analysis of high-dimensional data with sparse regression methods. The lasso was proposed that induces sparsity using an $L1$-norm penalty on all coefficients [4]. With the introduction of computationally efficient algorithms [5, 6], the lasso has since become a widely-applied variable selection method. Other methods for sparse regression include the smoothly clipped absolute deviation (SCAD) [7], adaptive lasso [8], Dantzig selector [9], etc. However, these methods were not designed for applications with mandatory covariates. An ad hoc approach is often employed where the response is regressed on mandatory covariates without penalization, as if in an ordinary least squares (OLS), while penalized regression is applied upon the optional variables, independently of the mandatory ones, to achieve variable selection. However, standard statistical principle advocates the consideration of all covariates simultaneously in order to account for complex dependencies among covariates. By penalizing coefficients disparately on some of the variables while not on others, this approach can yield both poor prediction accuracy and unreliable selection of optional variables. As mandatory covariates are routinely encountered in genomic-data analysis, it is of interest to develop a principled approach towards sparse regression with mandatory covariates. In this article, we consider the problem of efficient estimation of coefficients of mandatory covariates and simultaneous variable selection of optional variables.

Cancer arises as a disorder of the cell life cycle that leads to excessive cell proliferation and poor differentiation. Pathologists often use grading systems to measure the degree of cell differentiation in tumors [10, 11]. Tumor grade is one of the most important indicators for clinicians to guide treatment options and make prognosis for patients [12]. Histologic grade of breast cancer is representative of its aggressive potential [13]. Cancer cells with higher grades tend to be more aggressive and require quite different treatment strategies than those with lower grades. Due to the importance of tumor grade as an essential measure in clinical prognosis, treatment and of the survival of breast cancer patients, understanding genetic factors that may be predictive of tumor grade has become a desideratum of current research in breast cancers. In this article, we will propose a principled method to identify genes which may affect tumor grade while accounting for their clinical phenotypes such as age at diagnosis, p53 sequence mutation status, etc. by incorporating them as mandatory covariates.

We propose the *rid*ge-*l*asso hybrid *e*stimator (ridle), a novel penalized regression procedure that can simultaneously estimate coefficients of mandatory covariates while allowing selection for others. The ridle employs the $L2$-norm penalty to estimate mandatory coefficients and the $L1$-norm penalty to perform variable selection on the optional set. The $L2$-norm penalty has been successfully employed in ridge regression to efficiently estimate coefficients under a spectrum of dependency structures [14–16]. In this article, we provide theoretical, simulation, and real-data analysis to suggest the ridle as an efficient method for sparse regression with mandatory covariates. In particular, we will show that the ridle can achieve improved prediction accuracy and variable selection under commonly encountered scenarios when (1) the mandatory covariates are highly correlated among themselves or (2) the mandatory variables are correlated with the optional ones.

The rest of the article is presented as follows: "Methods" section introduces the ridle procedure, where an efficient algorithm is introduced and theoretical results are provided to suggest the efficacy of the ridle for sparse regression under mandatory predictors. "Results" section evaluates our method on simulated data. Further, we apply our method to a gene selection analysis of microarray data, where we identified more genes in breast-cancer related pathways with the ridle. Additionally, the ridle is the only method that identified two genes AREG and TRPM4 from the ErbB signaling pathway and ion-channel family, respectively, which are known to be related to cancer. Further discussions are provided in "Discussion" section, and we conclude with "Conclusions" section.

Methods

The Ridle

Consider the linear regression model,

$$\mathbf{y} = \mathbf{X}\beta^0 + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is an n-dimensional vector of random responses, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d)$ is the $n \times d$ design matrix, $\beta^0 = (\beta_1^0, \beta_2^0, \ldots, \beta_d^0)$ is a vector of regression parameters, and $\boldsymbol{\epsilon}$ is an n-dimensional vector of independent and identically distributed (i.i.d.) random variables with mean 0 and variance σ^2. We further assume that the response is centered and each predictor $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$ is standardized to have variance 1.

We define the *ridle* as,

$$\hat{\beta}(\lambda_1,\lambda_2) = \arg\min_{\beta} \left\{ ||\mathbf{y}-\mathbf{X}\beta||^2 + \lambda_1 \sum_{j\in\mathcal{O}} |\beta_j| + \lambda_2 \sum_{j\in\mathcal{M}} \beta_j^2 \right\}, \tag{1}$$

where $\mathcal{O}$ and $\mathcal{M}$ are non-intersecting subsets of the indices $\mathcal{I} = \{1, 2, \ldots, d\}$ such that $\mathcal{O} \cup \mathcal{M} = \mathcal{I}$. Subsets $\mathcal{O}$ and $\mathcal{M}$ comprise, respectively, indices of optional and mandatory variables. The ridle penalizes coefficients in $\mathcal{O}$ by the $L1$-norm penalty and coefficients in $\mathcal{M}$ by the $L2$-norm penalty. It allows variable selection, as in the lasso, for predictors in $\mathcal{O}$ and estimation without selection, as in the ridge, for predictors in $\mathcal{M}$. If λ_2 is equal to 0, the ridle is equivalent to thresholding the coefficients of some predictors for variable selection and estimating the rest without penalization. As the lasso penalty is applied to optional variables while no penalization is imposed on coefficients in $\mathcal{M}$, we call the special case of the ridle when $\lambda_2 = 0$ as the $\mathcal{M}$-unpenalized lasso.

For further insight, we examine the ridle estimator under two special situations.

Ridle estimator in two special cases

Orthogonal design case.

For $\mathbf{X}^T\mathbf{X}/n$ equal to the identity matrix $\mathbf{I}$, we can easily obtain the ridle solution in terms of the ordinary least squares estimates $\hat{\beta}_j(\text{ols})$,

$$\hat{\beta}_j = \begin{cases} \text{sign}(\hat{\beta}_j(\text{ols}))(|\hat{\beta}_j(\text{ols})| - \frac{\lambda_1}{2n})^+, & j \in \mathcal{O} \\ (1+\frac{\lambda_2}{n})^{-1}\hat{\beta}_j(\text{ols}), & j \in \mathcal{M} \end{cases} \tag{2}$$

where $(\cdot)^+$ denotes the positive part of the value, such that the expression is set to 0 for negative quantities. The ridle estimates equate to those of the lasso for $j \in \mathcal{O}$ and the ridge for $j \in \mathcal{M}$. When $\lambda_2 = 0$, the $\mathcal{M}$-unpenalized lasso estimates equate to those of the lasso for $j \in \mathcal{O}$ and the OLS for $j \in \mathcal{M}$. It is clear that, when the design matrix is orthogonal, the $L1$-norm and $L2$-norm penalties work independently to penalize coefficients with indices in $\mathcal{O}$ and $\mathcal{M}$, respectively. The situation is more involved when predictors are correlated.

Two-predictor case

Consider the case when $d = 2$. Let $\mathcal{O} = \{1\}$ and $\mathcal{M} = \{2\}$ for the ridle estimates. Figure 1 presents the penalty contours of the lasso, ridle, and ridge estimators. The ellipses centered at the OLS solutions are the contours of the quadratic loss function,

$$(\beta - \hat{\beta}(\text{ols}))^T\mathbf{X}^T\mathbf{X}(\beta - \hat{\beta}(\text{ols})),$$

plus a constant. With standardized predictors, these elliptical contours are at a $\pm$ 45°angle to the horizontal axes. Solutions occur when the ellipses first contact the penalty contours.

In Fig. 1a, we obtain the lasso solution as an ellipse hits a corner of the lasso penalty contour, setting β_1 to 0. In Fig. 1c, we see that the ridge penalty contour is circular, and an ellipse hitting the penalty contour gives nonzero estimates. The ridle penalty is described in Fig. 1b. It has both the characteristics of the lasso and ridge with an oval shape along the horizontal and sharp corners on the vertical axis. The ridle solution occurs when an ellipse centered on the OLS estimates hits a sharp corner on the vertical axis, yielding $\beta_1 = 0$ and a nonzero β_2. Thus, we see that the ridle may provide sparse solutions for coefficients in $\mathcal{O}$ while preserving non-sparsity for coefficients in $\mathcal{M}$.

When $d = 2$, we have the design matrix,

$$\mathbf{X}^T\mathbf{X} = n\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{3}$$

with pairwise correlation ρ. We can show that the ridle estimates are

$$\hat{\beta}_1 = s_1\Big(|\hat{\beta}_1(\text{ols})| + \frac{\rho\lambda_2}{n-n\rho^2}\theta_2 s_1\hat{\beta}_2(\text{ols}) - \theta_1\Big)^+,$$

$$\hat{\beta}_2 = \begin{cases} \theta_2\hat{\beta}_2(\text{ols}) + \frac{n\rho}{n+\lambda_2}s_1\theta_1, \\ \qquad\qquad \text{if } \theta_1 < |\hat{\beta}_1(\text{ols})| + \frac{\rho\lambda_2}{n-n\rho^2}\theta_2 s_1\hat{\beta}_2(\text{ols}), \\ \frac{n+\lambda_2-n\rho^2}{(n+\lambda_2)(1-\rho^2)}\theta_2\hat{\beta}_2(\text{ols}) + \frac{n\rho}{n+\lambda_2}\hat{\beta}_1(\text{ols}), \quad \text{otherwise,} \end{cases}$$

where $s_1 = \text{sign}(\hat{\beta}_1(\text{ols}))$, $\theta_1 = \lambda_1(n+\lambda_2)/(2n(n+\lambda_2-n\rho^2))$, and $\theta_2 = n(1-\rho^2)/(n+\lambda_2-n\rho^2)$. We see that the coefficient $\hat{\beta}_1$ can be thresholded to 0 with increasing $\theta_1(\lambda_1,\lambda_2,\rho)$ and $\theta_2(\lambda_2,\rho)$ functions to increase (temper) the thresholding of $\hat{\beta}_1$ when $\rho s_1\hat{\beta}_2(\text{ols})$ is negative (positive). On the other hand, $\hat{\beta}_2$ converges to a

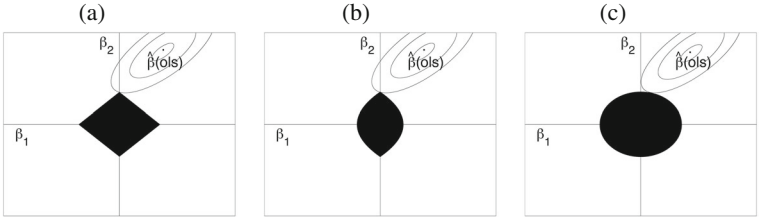

Fig. 1 Penalty contour when $d = 2$ for (**a**) lasso, (**b**) ridle, and (**c**) ridge regressions

weighted average of $\hat{\beta}_1(\text{ols})$ and $\hat{\beta}_2(\text{ols})$ without necessarily thesholding it to 0 as θ_1 increases to $|\hat{\beta}_1(\text{ols})| + \rho\lambda_2\theta_2 s_1\hat{\beta}_2(\text{ols})/(n - n\rho^2)$.

In the special case when $\lambda_2 = 0$, the ridle is reduced to the $\mathcal{M}$-unpenalized lasso, with estimates $\hat{\beta}_1(\mathcal{M}\text{-unpenalized lasso}) = s_1(|\hat{\beta}_1(\text{ols})| - \theta_1)^+$, $\hat{\beta}_2(\mathcal{M}\text{-unpenalized lasso}) = \hat{\beta}_2(\text{ols}) + s_1\rho\theta_1$ if $\theta_1 < |\hat{\beta}_1(\text{ols})|$, and $\hat{\beta}_2(\mathcal{M}\text{-unpenalized lasso}) = \hat{\beta}_2(\text{ols}) + \rho\hat{\beta}_1(\text{ols})$ otherwise. Under multicollinearity when ρ is large, the OLS estimates are known to have large variability. In this case, the ridge is often employed to improve prediction accuracy by regulating variances. Compared with the ridle, the $\mathcal{M}$-unpenalized lasso that imposes no penalization on mandatory coefficients can be less effective in tempering the effects of multicollinearity. For example, when $\rho = 1$ and θ_1 is large, the $\mathcal{M}$-unpenalized lasso estimate for β_2 is $\hat{\beta}_2(\text{ols}) + \hat{\beta}_1(\text{ols})$, such that the $\mathcal{M}$-unpenalized lasso can have larger prediction error than the OLS.

The lasso has the solution $\hat{\beta}_j(\text{lasso}) = s_j(|\hat{\beta}_j(\text{ols})| - \gamma)^+$ for $j = 1, 2$ and does not involve the correlation ρ when $d = 2$ [4]. In contrast, ridge coefficients tend to be averaged with increasing correlation. This property helps ridge to reduce variances of its estimates and improve prediction accuracy when data is multicollinear [16]. The ridle estimates $(\hat{\beta}_1, \hat{\beta}_2)$ are also defined in terms of weighted averages of $\hat{\beta}_1(\text{ols})$ and $\hat{\beta}_2(\text{ols})$ according to correlation ρ. In the following, we will show via theoretical studies how this property can improve variable selection for ridle.

Theoretical properties

In this section, we provide theoretical properties of the ridle estimator. These results are useful in providing a window to understanding the proposed method and a guide as to how the methods might perform in practice. Here, we use the sign-consistency approach [17] for theoretical derivations, which can provide results that are easy to interpret and relate to applications. More involved theoretical approaches, such as the asymptotic and non-asympotic oracle properties [7, 18], often rely on complex conditions that are difficult to interpret. Proofs for theoretical results in this section are provided in Additional file 1.

Without loss of generality, we assume that the true coefficients $\beta^0 = ((\beta^0_{(1)})^T, (\beta^0_{(2)})^T, (\beta^0_{(3)})^T)^T$ are partitioned such that $\beta^0_{(1)} = \{\beta^0_j : \beta^0_j \neq 0 \text{ and } j \in \mathcal{O}\}$, $\beta^0_{(2)} = \{\beta^0_j : \beta^0_j = 0 \text{ and } j \in \mathcal{O}\}$, and $\beta^0_{(3)} = \{\beta^0_j : j \in \mathcal{M}\}$. Let $\mathbf{C}^n = \mathbf{X}_n^T\mathbf{X}_n/n$ and $\tilde{\mathbf{C}}^n = \mathbf{C} + (\lambda_2/n)\mathbf{I}$. With the columns of $\mathbf{X}_n$ partitioned as β^0, $\mathbf{C}^n$ has the expression

$$\mathbf{C}^n = \begin{pmatrix} \mathbf{C}^n_{11} & \mathbf{C}^n_{12} & \mathbf{C}^n_{13} \\ \mathbf{C}^n_{21} & \mathbf{C}^n_{22} & \mathbf{C}^n_{23} \\ \mathbf{C}^n_{31} & \mathbf{C}^n_{32} & \mathbf{C}^n_{33} \end{pmatrix}. \quad (4)$$

We assume that

$$\mathbf{C}^n \to \mathbf{C}, \quad (5)$$

where $\mathbf{C}$ is a positive definite matrix,

$$\frac{1}{n}\max_{1\le i\le n}((\mathbf{x}^n_i)^T\mathbf{x}^n_i) \to 0, \quad (6)$$

and

$$\tilde{\mathbf{C}}^n_{33} \text{ and } (\mathbf{C}^n_{11} - \mathbf{C}^n_{13}(\tilde{\mathbf{C}}^n_{33})^{-1}\mathbf{C}^n_{31}) \text{ are invertible.} \quad (7)$$

Asymptotic normality of the Ridle

Theorem 1 *With (5) and (6), $\hat{\beta}(\lambda_1, \lambda_2)$ satisfies the following for $\lambda_1, \lambda_2 > 0$ such that $\lambda_1/\sqrt{n} \to c_1 < \infty$ and $\lambda_2/\sqrt{n} \to c_2$.*

$$\sqrt{n}(\hat{\beta} - \beta^0) \to_d \arg\min V(\mathbf{u}) \quad (8)$$

where

$$\begin{aligned} V(\mathbf{u}) = {} & \mathbf{u}^T\mathbf{C}\mathbf{u} - 2\mathbf{u}^T\mathbf{W} \\ & + c_1\sum_{j\in\mathcal{O}}(u_j sign(\beta^0_j)I(\beta^0_j \neq 0) + |u_j|I(\beta^0_j = 0)) \\ & + 2c_2\mathbf{u}^T_{(3)}\beta^0_{(3)} \end{aligned}$$

and $\mathbf{W} \sim N(\mathbf{0}, \sigma^2\mathbf{C})$.

Theorem 1 shows that the coefficients of mandatory covariates can contribute to the biasness of the ridle estimates. If the coefficients of variables in $\mathcal{M}$ are relatively large, then a small c_2 is required to keep the bias low. On the other hand, when coefficients of variables in $\mathcal{M}$ are small, a wider spectrum of values of c_2 can be chosen to improve prediction accuracy. Hence, we expect the ridle to perform the best when coefficients of mandatory variables tend to be small.

Variable selection consistency of the Ridle.

In this section, we provide sign consistency results for the ridle. We define the ridle estimates $\hat{\beta}(\lambda_1, \lambda_2)$ to be sign consistent if there exist $\lambda_1 = \lambda_1(n)$ and $\lambda_2 = \lambda_2(n)$ such that

$$\lim_{n\to\infty} P(\text{sign}(\hat{\beta}_{\mathcal{O}}(\lambda_1, \lambda_2)) = \text{sign}(\beta^0_{\mathcal{O}})) = 1. \quad (9)$$

Sign consistency, as a stronger condition, directly implies variable selection consistency.

With (5)-(7), we give the following conditions for sign consistency of the ridle estimator.

Sufficient condition: There exists $\eta > 0$ such that

$$|\mathbf{D}^n\text{sign}(\beta^0_{(1)}) - \frac{2\lambda_2}{\lambda_1}(\mathbf{D}^n\mathbf{C}^n_{13} - \mathbf{C}^n_{23})(\tilde{\mathbf{C}}^n_{33})^{-1}\beta^0_{(3)}| \le \mathbf{1} - \eta, \quad (10)$$

where $\mathbf{D}^n = (\mathbf{C}^n_{21} - \mathbf{C}^n_{23}(\tilde{\mathbf{C}}^n_{33})^{-1}\mathbf{C}^n_{31})(\mathbf{C}^n_{11} - \mathbf{C}^n_{13}(\tilde{\mathbf{C}}^n_{33})^{-1}\mathbf{C}^n_{31})^{-1}$.

Necessary condition:

$$|\mathbf{D}^n \text{sign}(\beta^0_{(1)}) - \frac{2\lambda_2}{\lambda_1}(\mathbf{D}^n\mathbf{C}^n_{13} - \mathbf{C}^n_{23})(\tilde{\mathbf{C}}^n_{33})^{-1}\beta^0_{(3)}| < \mathbf{1}. \quad (11)$$

Theorem 2 *Under (5)-(7), $\hat{\beta}(\lambda_1, \lambda_2)$ is sign consistent if condition (10) holds for $\lambda_1, \lambda_2 > 0$ such that $\lambda_1/n \to 0$, $\lambda_1/\sqrt{n} \to \infty$, and $\lambda_2/\lambda_1 \to c < \infty$.*

Theorem 3 *Under (5)-(7), $\hat{\beta}(\lambda_1, \lambda_2)$ is sign consistent only if condition (11) holds for $\lambda_1, \lambda_2 > 0$ such that $\lambda_2/n \to 0$.*

Remark 1 *Let $\mathbf{C}^n_{12} = \mathbf{C}^n_{23} = \mathbf{0}$. Then conditions (10) and (11) are satisfied with left-hand sides equal to $\mathbf{0}$. Thus, the ridle estimator is sign consistent when predictors with nonzero coefficients are unrelated with predictors with zero coefficients.*

Remark 2 *Suppose $\mathbf{C}^n_{13} = \mathbf{C}^n_{23} = \mathbf{0}$. Then, conditions (10) and (11) become*

$$|\mathbf{C}^n_{21}(\mathbf{C}^n_{11})^{-1}sign(\beta^0_{(1)})| < 1 - \eta \quad and$$
$$|\mathbf{C}^n_{21}(\mathbf{C}^n_{11})^{-1}sign(\beta^0_{(1)})| < 1,$$

respectively. This is equivalent to the Irrepresentable conditions of [17] *for lasso sign consistency. Thus, when variables in $\mathcal{M}$ are uncorrelated with those in $\mathcal{O}$, sign consistency for the ridle is equivalent to that of the lasso for predictors in $\mathcal{O}$.*

Remark 3 *Consider performing the lasso on predictors in both $\mathcal{O}$ and $\mathcal{M}$. The following Irrepresentable conditions are derived in Zhao and Yu (2006) for the lasso* [17],

$$\textit{(sufficient)} \qquad |\mathbf{D}^n_1 sign(\beta^0_{(1)}) + \mathbf{D}^n_2 sign(\beta^0_{(3)})| < 1 - \eta \quad (12)$$

$$\textit{(necessary)} \qquad |\mathbf{D}^n_1 sign(\beta^0_{(1)}) + \mathbf{D}^n_2 sign(\beta^0_{(3)})| < 1 \quad (13)$$

where $\mathbf{D}^n_1 = \mathbf{C}^n_{21} - \mathbf{C}^n_{23}(\mathbf{C}^n_{33})^{-1}\mathbf{C}^n_{31})(\mathbf{C}^n_{11} - \mathbf{C}^n_{13}(\mathbf{C}^n_{33})^{-1}\mathbf{C}^n_{31})^{-1}$ and $\mathbf{D}^n_2 = (\mathbf{C}^n_{23} - \mathbf{C}^n_{21}(\mathbf{C}^n_{11})^{-1}\mathbf{C}^n_{13})(\mathbf{C}^n_{33} - \mathbf{C}^n_{31}(\mathbf{C}^n_{11})^{-1}\mathbf{C}^n_{13})^{-1}$. Compared with the lasso, the ridle conditions in (10) and (11) involve the parameter λ_2 that can allow it to more easily satisfy sign consistency conditions with suitable choices of λ_2. Consider a toy example to better understand the general results through a simple scenario. If $\mathbf{C}^n_{23} \neq \mathbf{0}$ and $\mathbf{C}^n_{12} = \mathbf{C}^n_{13} = \mathbf{0}$, then the left-hand sides of conditions (10) and (11) for the ridle become

$$|2(\lambda_2/\lambda_1)\mathbf{C}^n_{23}(\tilde{\mathbf{C}^n_{33}})^{-1}\beta^0_{(3)}|,$$

whereas the left-hand sides of conditions (12) and (13) for the lasso are

$$|\mathbf{C}^n_{23}(\mathbf{C}^n_{33})^{-1}sign(\beta^0_{(3)})|.$$

Here, the ridle is sign consistent for λ_2/λ_1 sufficiently small, but the lasso is not if the elements of $\mathbf{C}^n_{23}$ are large or $\mathbf{C}^n_{33}$ is nearly singular. Thus, we expect that the ridle may perform better than lasso using predictors from both $\mathcal{O}$ and $\mathcal{M}$ in terms of model selection if predictors in $\mathcal{M}$ are correlated with the irrelevant predictors in $\mathcal{O}$ or the predictors in $\mathcal{M}$ are highly correlated among themselves.

Remark 4 *When $\lambda_2 = 0$, the ridle is reduced to the $\mathcal{M}$-unpenalized lasso. In this case, as no penalization is involved on mandatory coefficients, sign consistency conditions can be trivially obtained as*

$$\left|\mathbf{D}^n_1 sign\left(\beta^0_{(1)}\right)\right| < 1 - \eta \qquad and \qquad \left|\mathbf{D}^n_1 sign\left(\beta^0_{(1)}\right)\right| < 1, \quad (14)$$

where D^n_1 is as defined following conditions (12) and (13) for the lasso. Compared with the $\mathcal{M}$-unpenalized lasso, the ridle composes of an additional offsetting factor $(2\lambda_2/\lambda_1)(\mathbf{D}^n\mathbf{C}^n_{13} - \mathbf{C}^n_{23})(\tilde{\mathbf{C}}^n_{33})^{-1}\beta^0_{(3)}$ in conditions (10) and (11) that allows sign consistency inequalities to be more easily satisfied. For example, if $\mathbf{C}^n_{23} = \mathbf{0}$, $\mathbf{C}^n_{12} \neq \mathbf{0}$, and $\mathbf{C}^n_{13} \neq \mathbf{0}$, then the left-hand sides of conditions (10) and (11) for the ridle become

$$\left|\mathbf{D}^n sign\left(\beta^0_{(1)}\right) - 2\left(\lambda_2/\lambda_1\right)\mathbf{D}^n\mathbf{C}^n_{13}\left(\tilde{\mathbf{C}^n_{33}}\right)^{-1}\beta^0_{(3)}\right|.$$

In this case, if $|\mathbf{D}^n_1 sign(\beta^0_{(1)})| \geq 1$, sign consistency conditions for the $\mathcal{M}$-unpenalized lasso in (14) are violated, whereas sign consistency conditions for the ridle may still be satisfied with suitably chosen λ_2.

Remark 5 *When the mandatory covariates are irrelevant, $\beta^0_{(3)} = \mathbf{0}$, the offsetting terms in (10) and (11) for ridle sign consistency would vanish. Indeed, with $\lambda_2/n \to 0$, ridle sign consistency conditions are equivalent to those of both the lasso and $\mathcal{M}$-unpenalized lasso. However, this does not mean that the methods will perform similarly under finite samples. We will examine their finite-sample performances in the Analysis of Simulated Data in the "Results" section.*

Efficient algorithm

We provide an efficient algorithm for computing the ridle. Programming code, written in Fortran, and its R-language wrapper for the algorithm described in this section are freely available online at http://sites.google.com/site/zhongyindaye/software.

The ridle (1) minimizes over an objective function with convex and separable penalties. This allows us to employ the coordinate descent strategy [19–21] to compute for the ridle. In the coordinate descent, we update first for all coefficients of mandatory variables $\beta_{\mathcal{M}}$ and, then, each optional coefficient $\beta_j \in \mathcal{O}$, one at a time. This is iterated till practical convergence is reached. The algorithm is further sped up by iterating only through the mandatory and

active set till convergence before updating all variables. We provide the coordinate descent updating equations as the following,

$$\beta_{\mathcal{M}} \leftarrow \left(\mathbf{X}_{\mathcal{M}}^T\mathbf{X}_{\mathcal{M}} + \lambda_2\mathbf{I}\right)^{-1}\mathbf{X}_{\mathcal{M}}^T\left(\mathbf{y} - \mathbf{X}_{\mathcal{O}}\beta_{\mathcal{O}}\right) \quad (15)$$

$$\beta_j \leftarrow \frac{s_j}{\|\mathbf{x}_j\|^2}\left(\left|\mathbf{x}_j^T\left(\mathbf{y} - \sum_{k\neq j}\mathbf{x}_k\beta_k\right)\right| - \frac{\lambda_1}{2}\right)_+ \text{ for } j \in \mathcal{O}, \quad (16)$$

where $s_j = \text{sign}(\mathbf{x}_j^T(\mathbf{y} - \sum_{k\neq j}\mathbf{x}_k\beta_k))$. Maximum value for λ_1 is $\lambda_1^{max} = 2\max_{j\in\mathcal{O}}|\mathbf{y} - \mathbf{X}_{\mathcal{M}}\hat{\beta}_{\mathcal{M},0}|$, where $\hat{\beta}_{\mathcal{M},0} = (\mathbf{X}_{\mathcal{M}}^T\mathbf{X}_{\mathcal{M}} + \lambda_2\mathbf{I})^{-1}\mathbf{X}_{\mathcal{M}}^T\mathbf{y}$ are initial estimates for coefficients of mandatory covariates. The matrix inverse in (15) can be computed efficiently by taking the inverse of individual eigenvalues added to λ_2 after an initial singular value decomposition of $\mathbf{X}_{\mathcal{M}}^T\mathbf{X}_{\mathcal{M}}$.

Results

Analysis of simulated data

We evaluate the performances of the ridle via simulation studies. We examine effects of having different magnitudes of coefficients, correlations between mandatory and irrelevant predictors, and degrees of multicollinearity among mandatory covariates. We compare the ridle to the ridge, lasso, elastic net, and the lasso and elastic net without penalization on the mandatory covariates. We use the R package *glmnet 2.0* to compute for the lasso and elastic net, where the penalization on mandatory covariates is specified using the `penalty.factor` option.

In each example, we simulate 200 times from the true model, $\mathbf{y} = \mathbf{X}\beta + \sigma\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$. We use $n = 50$ number of observations and $p = 250$ predictors. Tuning parameters are estimated using 5-fold cross-validation. We measure prediction accuracy using the relative prediction error, $rpe = (\hat{\beta} - \beta)^T\boldsymbol{\Sigma}(\hat{\beta} - \beta)/\sigma^2$, where $\boldsymbol{\Sigma}$ is the population covariance matrix. Further, we examine variable selection performances using sensitivity, specificity, and g-measure. Sensitivity and specificity are, respectively, the marginal proportions of selecting relevant variables and discarding irrelevant variables correctly. In other words, sensitivity is the proportion of true positives among all relevant variables, whereas specificity is the proportion of true negatives among all irrelevant variables. The false positive rate is equal to 1 - specificity. As proportions, sensitivity and specificity allow intuitive comparisons across simulation settings with varying numbers of relevant and irrelevant variables in high-dimensional variable selection. We examine overall variable selection performances using the g-measure, $\sqrt{sensitivity * specificity}$. A g-measure close to 1 indicates accurate variable selection, whereas a g-measure close to 0 implies that few relevant variables or too many irrelevant variables are selected, or both. In Tables 1, 2, 3 and 4, we report medians and bootstrapped standard deviations of medians out of 500 re-samplings, in parentheses. Further, we boldface, as top measurements, the smallest rpe and two largest g-measures in each case.

Example 1 (Effect of signal strengths)

This example has $\beta_j = \beta_0$ for $j \in \{1, \ldots, 10, 21, \ldots, 30\}$ and $\beta_j = 0$ otherwise. Predictors are generated from $\mathbf{X} \sim \mathbf{N}(\mathbf{0}, \Sigma)$ where $\Sigma_{ij} = 0.5^{|i-j|}$. $\sigma = 3$. We assume the mandatory covariates to be comprised of the relevant variables so that $\mathcal{M} = \{1, \ldots, 10, 21, \ldots, 30\}$.

Table 1 displays prediction accuracy and variable selection performances for this example. First of all, by utilizing a priori information on mandatory covariates, the ridle has significantly smaller *rpe*'s than those of the ridge, lasso and elastic net with or without penalization on mandantory covariates. Additionally, the ridle has larger g-measure than those of the lasso and elastic net and similar g-measure with the mandantory-unpenalized lasso and elastic-net method. Sensitivity for the lasso and elastic net decreases dramatically as the signal strength weakens or β_0 becomes smaller. On the other hand, specificity for the lasso decreases while increasing for elastic net when β_0 becomes larger. Furthermore, the lasso and elastic net without penalization on mandantory variables outperforms the lasso and elastic net with penaliztion on the mandantory variables in terms of both prediction accuracy and variable selection. These suggest that, even though the elastic net does a better job than the lasso in terms of prediction accuracy, both methods may not be able to distinguish well between mandatory and irrelevant variables, and incorporating a priori knowledge on mandatory covariates can yield significant improvements.

Example 2 (Effect of correlation between mandatory and irrelevant predictors)

In this example, we have $\beta_j = 2$ for $j \in \{2k : k = 1, \ldots, 10\}$, $\beta_j = 1.5$ for $j \in \{2k : k = 11, \ldots, 20\}$, and $\beta_j = 0$ otherwise. Predictors are generated from $\mathbf{X} \sim \mathbf{N}(\mathbf{0}, \Sigma)$ where each element $\Sigma_{ij} = \rho_0^{|i-j|}$. Thus, relevant predictors are interspersed with irrelevant ones, to which they are correlated. Further, we assume $\mathcal{M} = \{2k : k = 11, \ldots, 20\}$ and $\sigma = 6$. $\Sigma_{ij} = \rho_0^{|i-j|}$ presents an autocorrelated dependence structure, such that a variable $\mathbf{x}_j$ has a correlation of ρ_0 with its immediate neighbors $\mathbf{x}_{j-1}$ and $\mathbf{x}_{j+1}$ for $1 < j < p$. When ρ_0 is large, each variable is highly correlated with its immediate neighbors, resulting in multicollinearity.

Table 2 presents prediction accuracy and variable selection performances for this example. The ridle performs the best in terms of *rpe*'s. When ρ_0 is large at 0.75, mandatory covariates are strongly correlated with some of the optional variables, and the $\mathcal{M}$-unpenalized lasso performs the worst in terms of prediction accuracy,

Table 1 Simulation example 1: effect of signal strengths

	Method	*rpe*	*g*-measure	Sensitivity	Specificity
$\beta_0 = 0.5$	Ridge	1.008 (0.009)			
	Lasso	1.004 (0.018)	0.582 (0.009)	0.350 (0.018)	0.957 (0.006)
	Elastic net	0.923 (0.020)	0.676 (0.007)	0.600 (0.041)	0.848 (0.023)
	$\mathcal{M}$-unpenalized lasso	0.675 (0.028)	**1.000** (0.000)	1.000 (0.000)	1.000 (0.000)
	$\mathcal{M}$-unpenalized elastic net	0.697 (0.026)	**1.000** (0.001)	1.000 (0.000)	1.000 (0.002)
	Ridle	**0.281** (0.016)	**0.998** (0.001)	1.000 (0.000)	0.996 (0.002)
$\beta_0 = 1.5$	Ridge	6.549 (0.056)			
	Lasso	3.300 (0.083)	0.839 (0.005)	0.750 (0.017)	0.926 (0.003)
	elastic net	3.230 (0.118)	0.853 (0.004)	0.900 (0.008)	0.850 (0.005)
	$\mathcal{M}$-unpenalized lasso	0.691 (0.023)	**1.000** (0.000)	1.000 (0.000)	1.000 (0.000)
	$\mathcal{M}$-unpenalized elastic net	0.701(0.028)	**1.000** (0.001)	1.000 (0.000)	1.000 (0.001)
	Ridle	**0.473** (0.014)	**0.998** (0.001)	1.000 (0.000)	0.996 (0.002)
$\beta_0 = 3$	Ridge	24.559 (0.317)			
	Lasso	8.074 (0.433)	0.908 (0.005)	0.900 (0.013)	0.935 (0.002)
	Elastic net	6.735 (0.339)	0.903 (0.002)	0.950 (0.013)	0.852 (0.003)
	$\mathcal{M}$-unpenalized lasso	0.676 (0.032)	**1.000** (0.000)	1.000 (0.000)	1.000 (0.000)
	$\mathcal{M}$-unpenalized elastic net	0.725 (0.030)	**1.000** (0.000)	1.000 (0.000)	1.000 (0.001)
	Ridle	**0.605** (0.025)	**0.998** (0.001)	1.000 (0.000)	0.996 (0.002)

The $\mathcal{M}$-unpenalized lasso and $\mathcal{M}$-unpenalized elastic net were performed without penalization on the mandatory covariates
$n = 50, p = 250, |\mathcal{M}| = 20$. The smallest *rpe* and largest two *g*-measures are boldfaced

Table 2 Simulation example 2: effect of correlation between mandatory and irrelevant predictors

	Method	*rpe*	*g*-measure	Sensitivity ($\mathcal{M}$)	Sensitivity ($\mathcal{O}$)	Specificity ($\mathcal{O}$)
$\rho_0 = 0.25$	Ridge	1.671 (0.012)				
	Lasso	1.911 (0.022)	0.383 (0.034)	0.100 (0.032)	0.200 (0.028)	0.975 (0.008)
	Elastic net	1.744 (0.019)	0.585 (0.015)	0.400 (0.054)	0.600 (0.050)	0.835 (0.036)
	$\mathcal{M}$-unpenalized lasso	1.741 (0.028)	0.742 (0.012)	1.000 (0.000)	0.200 (0.037)	0.938 (0.003)
	$\mathcal{M}$-unpenalized elastic net	1.657 (0.017)	**0.757** (0.008)	1.000 (0.000)	0.500 (0.064)	0.833 (0.022)
	Ridle	**1.492** (0.031)	**0.773** (0.006)	1.000 (0.000)	0.200 (0.048)	0.931 (0.006)
$\rho_0 = 0.5$	Ridge	1.807 (0.014)				
	Lasso	2.045 (0.035)	0.571 (0.013)	0.300 (0.046)	0.400 (0.039)	0.925 (0.007)
	Elastic net	1.773 (0.034)	0.667 (0.008)	0.600 (0.014)	0.800 (0.048)	0.756 (0.020)
	$\mathcal{M}$-unpenalized lasso	1.922 (0.044)	0.794 (0.003)	1.000 (0.000)	0.400 (0.047)	0.929 (0.004)
	$\mathcal{M}$-unpenalized elastic net	1.729 (0.040)	**0.796** (0.007)	1.000 (0.000)	0.700 (0.048)	0.785 (0.022)
	Ridle	**1.438** (0.057)	**0.852** (0.006)	1.000 (0.000)	0.600 (0.049)	0.900 (0.004)
$\rho_0 = 0.75$	Ridge	1.564 (0.022)				
	Lasso	1.365 (0.029)	0.684 (0.008)	0.400 (0.032)	0.600 (0.012)	0.900 (0.003)
	Elastic net	1.237 (0.030)	0.745 (0.005)	0.700 (0.048)	0.900 (0.011)	0.775 (0.014)
	$\mathcal{M}$-unpenalized lasso	1.423 (0.037)	0.839 (0.005)	1.000 (0.000)	0.700 (0.026)	0.904 (0.006)
	$\mathcal{M}$-unpenalized elastic net	1.310 (0.041)	**0.847** (0.005)	1.000 (0.000)	0.800 (0.012)	0.840 (0.008)
	Ridle	**0.886** (0.029)	**0.875** (0.003)	1.000 (0.000)	0.700 (0.038)	0.908 (0.003)

The $\mathcal{M}$-unpenalized lasso and $\mathcal{M}$-unpenalized elastic net were performed without penalization on the mandatory covariates. *g*-measure is estimated from all predictors. Sensitivity ($\mathcal{M}$) is computed in terms of the mandatory variables only, whereas sensitivity ($\mathcal{O}$) and specificity ($\mathcal{O}$) are computed in terms of the optional variables only
$n = 50, p = 250, |\mathcal{M}| = 10$. The smallest *rpe* and largest two *g*-measures are boldfaced

Table 3 Simulation example 3: effect of multicollinearity among mandatory covariates

	Method	*rpe*	*g*-measure	Sensitivity ($\mathcal{M}$)	Sensitivity ($\mathcal{O}$)	Specificity ($\mathcal{O}$)
$\rho = 0.75$	Ridge	6.353 (0.022)				
	Lasso	4.649 (0.167)	0.802 (0.011)	0.800 (0.000)	0.700 (0.048)	0.908 (0.004)
	Elastic net	4.410 (0.128)	0.804 (0.005)	1.000 (0.009)	0.700 (0.006)	0.858 (0.006)
	$\mathcal{M}$-unpenalized lasso	4.776 (0.260)	**0.829** (0.005)	1.000 (0.000)	0.700 (0.031)	0.902 (0.007)
	$\mathcal{M}$-unpenalized elastic net	5.402 (0.190)	0.823 (0.006)	1.000 (0.000)	0.700 (0.013)	0.871 (0.009)
	Ridle	**2.699** (0.152)	**0.893** (0.007)	1.000 (0.000)	0.900 (0.048)	0.904 (0.004)
$\rho = 0.9$	Ridge	6.270 (0.026)				
	Lasso	4.914 (0.148)	0.784 (0.010)	0.600 (0.089)	0.700 (0.036)	0.908 (0.004)
	Elastic net	4.336 (0.135)	0.816 (0.005)	0.800 (0.092)	0.700 (0.018)	0.867 (0.008)
	$\mathcal{M}$-unpenalized lasso	6.992 (0.337)	**0.828** (0.008)	1.000 (0.000)	0.700 (0.031)	0.902 (0.006)
	$\mathcal{M}$-unpenalized elastic net	7.245 (0.237)	0.827 (0.005)	1.000 (0.000)	0.700 (0.045)	0.860 (0.011)
	Ridle	**3.000** (0.214)	**0.890** (0.006)	1.000 (0.000)	0.800 (0.045)	0.900 (0.004)
$\rho = 0.99$	Ridge	6.231 (0.031)				
	Lasso	7.322 (0.200)	0.745 (0.005)	0.400 (0.000)	0.700 (0.000)	0.913 (0.003)
	Elastic net	5.003 (0.155)	0.804 (0.006)	0.800 (0.049)	0.700 (0.019)	0.883 (0.006)
	$\mathcal{M}$-unpenalized lasso	36.214 (2.064)	0.824 (0.006)	1.000 (0.000)	0.700 (0.046)	0.904 (0.005)
	$\mathcal{M}$-unpenalized elastic net	33.583 (2.197)	**0.830** (0.004)	1.000 (0.010)	0.700 (0.045)	0.867 (0.010)
	Ridle	**4.193** (0.343)	**0.890** (0.005)	1.000 (0.000)	0.800 (0.029)	0.904 (0.004)

The $\mathcal{M}$-unpenalized lasso and $\mathcal{M}$-unpenalized elastic net were performed without penalization on the mandatory covariates. *g*-measure is estimated from all predictors. Sensitivity ($\mathcal{M}$) is computed in terms of the mandatory variables only, whereas sensitivity ($\mathcal{O}$) and specificity ($\mathcal{O}$) are computed in terms of the optional variables only $n = 50, p = 250, |\mathcal{M}| = 5$. The smallest *rpe* and largest two *g*-measures are boldfaced

except that of the ridge. This corroborates comments in *Two-Predictor Case* of "Methods" section that suggest the $\mathcal{M}$-unpenalized lasso can have large prediction errors under multicollinearity. Further, the ridle performs the best in terms of *g*-measures for overall variable selection in all scenarios.

Example 3 (Effect of multicollinearity among mandatory covariates)

Here, we have $\beta_j = 3$ for $j \in \{1, \ldots, 5\}$, $\beta_j = 1.5$ for $j \in \{6, \ldots, 10\}$, $\beta_j = 2$ for $j \in \{16, \ldots, 20\}$, and $\beta_j = 0$ otherwise. We set $\sigma = 3$ and assume $\mathcal{M} = \{16, \ldots, 20\}$. Let $Z \sim N(0, 1)$ and $\epsilon_{\mathbf{x}} \sim \mathbf{N}(\mathbf{0}, \mathbf{1})$. We generate predictors as $\mathbf{x_j} = \mathbf{Z} + \sqrt{(\mathbf{1} - \rho)/\rho}\,\epsilon$ for $j \in \mathcal{M}$ and $\mathbf{x_j} \sim \mathbf{N}(\mathbf{0}, \mathbf{1})$ otherwise. This creates correlations of ρ among the mandatory covariates.

In Table 3, we see that sensitivity ($\mathcal{M}$) decreases for the lasso and elastic net as ρ increases. Additionally, the lasso and elastic net without penalization on mandatory variables have identical sensitivity ($\mathcal{M}$) with the ridle. Furthermore, prediction error for the lasso without penalization on mandatory covariates increases dramatically as ρ increases, whereas the ridle has the lowest *rpe*'s. This corroborates Remark 3 of *Variable Selection Consistency of the Ridle* in "Methods" section, which suggests that the ridle may outperform the lasso when mandatory variables are highly correlated among themselves.

Example 4 (Mandatory covariates are irrelevant)

We repeat the simulation setting from example 2, but with the mandatory covariates defined as $\mathcal{M} = \{2k - 1 : k = 1, \ldots, 10\}$. In this case, the mandatory covariates are irrelevant.

Table 4 presents prediction accuracy and variable selection performances for this scenario. The ridle underperforms the elastic net but outperforms all other variable selection methods in terms of prediction accuracy. Indeed, the ridle has significantly smaller *rpe*'s compared with the $\mathcal{M}$-unpenalized lasso and $\mathcal{M}$-unpenalized elastic net. Moreover, the ridle underperforms both the lasso and elastic net in terms of *g*-measures for overall variable selection. However, ridle outperforms the $\mathcal{M}$-unpenalized lasso and $\mathcal{M}$-unpenalized elastic net at $\rho_0 = 0.5$ and $\rho_0 = 0.75$, when the irrelevant mandatory covariates are moderately and highly correlated, respectively, with some of the relevant optional variables. These suggest that the ridle, although not designed to exclude mandatory covariates when they are irrelevant, can be more advantageous than related methods that include mandatory covariates,

Table 4 Simulation example 4: mandatory covariates are irrelevant

	Method	*rpe*	*g*-measure	Specificity ($\mathcal{M}$)	Sensitivity ($\mathcal{O}$)	Specificity ($\mathcal{O}$)
$\rho_0 = 0.25$	Ridge	**1.671** (0.012)				
	Lasso	1.911 (0.022)	**0.383** (0.034)	1.000 (0.000)	0.200 (0.028)	0.975 (0.008)
	Elastic net	1.744 (0.019)	**0.585** (0.015)	0.600 (0.053)	0.600 (0.050)	0.835 (0.036)
	$\mathcal{M}$-unpenalized lasso	2.357 (0.032)	0.215 (0.103)	0.000 (0.000)	0.050 (0.024)	0.995 (0.003)
	$\mathcal{M}$-unpenalized elastic net	2.210 (0.034)	0.308 (0.054)	0.000 (0.000)	0.525 (0.065)	0.732 (0.057)
	Ridle	1.854 (0.012)	0.309 (0.029)	0.000 (0.000)	0.100 (0.024)	0.982 (0.005)
$\rho_0 = 0.5$	Ridge	1.807 (0.014)				
	Lasso	2.045 (0.035)	**0.571** (0.013)	0.800 (0.006)	0.400 (0.039)	0.925 (0.007)
	Elastic net	**1.773** (0.034)	**0.667** (0.008)	0.500 (0.048)	0.800 (0.048)	0.756 (0.020)
	$\mathcal{M}$-unpenalized lasso	2.242 (0.023)	0.299 (0.035)	0.000 (0.000)	0.100 (0.021)	0.982 (0.004)
	$\mathcal{M}$-unpenalized elastic net	2.080 (0.028)	0.305 (0.094)	0.000 (0.000)	0.550 (0.072)	0.700 (0.079)
	Ridle	1.801 (0.039)	0.528 (0.032)	0.000 (0.000)	0.300 (0.038)	0.943 (0.005)
$\rho_0 = 0.75$	Ridge	1.564 (0.022)				
	Lasso	1.365 (0.029)	**0.684** (0.008)	0.700 (0.041)	0.600 (0.012)	0.900 (0.003)
	Elastic net	**1.237** (0.030)	**0.745** (0.005)	0.300 (0.046)	0.900 (0.011)	0.775 (0.014)
	$\mathcal{M}$-unpenalized lasso	1.747 (0.043)	0.428 (0.003)	0.000 (0.000)	0.200 (0.000)	0.964 (0.003)
	$\mathcal{M}$-unpenalized elastic net	1.662 (0.043)	0.514 (0.016)	0.000 (0.000)	0.350 (0.023)	0.900 (0.015)
	Ridle	1.253 (0.042)	0.596 (0.017)	0.000 (0.000)	0.400 (0.026)	0.945 (0.003)

The $\mathcal{M}$-unpenalized lasso and $\mathcal{M}$-unpenalized elastic net were performed without penalization on the mandatory covariates. *g*-measure is estimated from all predictors. specificity ($\mathcal{M}$) is computed in terms of the mandatory variables only, whereas sensitivity ($\mathcal{O}$) and specificity ($\mathcal{O}$) are computed in terms of the optional variables only $n = 50, p = 250, |\mathcal{M}| = 10$. The smallest *rpe* and largest two *g*-measures are boldfaced

as the ridle penalizes coefficients of irrelevant mandatory covariates towards, although not equal to, 0 with the ridge penalty.

Gene expression analysis on histologic grades of breast cancer

Histologic grades are an important determinant of the aggressive potential of breast cancers and are of practical importance in the assessment and choice of treatment options. In this section, we apply our proposed method on a microarray gene expression dataset to determine genes that may be predictive of breast tumor histologic grade [22]. In this experiment, 251 frozen tumor tissure were collected from primary breast cancer patients and more than 12,000 genes were assayed on 251 subjects. We removed 2 subjects with missing outcomes and performed our analysis with the remaining 249 observations. Clinicopathological variables, such as ER status, PgR status, age and tumor size, measured at diagnosis, were obtained from patient records. Histologic grades are based on the widely used Nottingham Histologic Score system for prognosis of breast cancer [23]. There are three factors that pathologists consider in this scoring system: cell differentiation, nuclear features and mitotic activity [24]. Considerations of these factors allow the Nottingham Prognostic Index (NPI) to provide comprehensive prognosis of breast cancers. The three factors are each assigned a score from 1–3 based on clinical observations. A tumor is assigned a score of 1, 2, or 3 for cell differentiation if >75%, 10%-75%, or <10% of tumor area form glandular structures, respectively. A tumor has a score of 1, 2, or 3 for nuclear features if nuclei have little increase in size, larger than normal breast epithelial cells, or prominent nucleoli with occasionally very large sizes, respectively. Further, breast tumors have scores of 1, 2, or 3 for mitotic activity if $\leq$7, 8–14, or $\geq$15 mitoses per 10 high power microscopic fields are observed, respectively. Overall tumor grades are obtained by summing the scores for the three factors. Breast tumors with total scores of 3–5, 6–7, and 8–9 are assigned with tumor grades 1 (low), 2 (intermediate), and 3 (high), respectively, that represent the aggressive potential of breast tumors. The higher the grade is, the more likely it will spread or become aggressive. This dataset is available at the NCBI Gene Expression Omnibus (GEO) repository with GEO accession: gse3294. We focused our analysis on 430 genes

from several well-known cancer-related pathways: PI3K [25, 26], p53 signaling [27–29], VEGF [30, 31], Hedgehog signaling [32, 33], ErbB signaling [34, 35], Ras signaling [36, 37] and Ion-channel family [38, 39].

Significant genes are selected as predictors of breast tumor grade along the 7 pathways by utilizing a sparse regression approach [40, 41]. In this strategy, tumor grade is regressed upon both the 4 clinicopathological covariates (ER status, PgR status, age and tumor size) and 430 gene expression levels, and significant predictor to tumor grade based on clinical covariates and genes are identified if they are retained in sparse regression analysis. We applied the ridle to perform variable selection on gene expression levels while conditioning on the 4 clinicopathological variables that we incorporated as mandatory covariates. We further compared our results with those from the ridge, lasso, elastic net, and lasso and elastic net without penalization on the 4 mandatory covariates.

Table 5 presents the numbers of genes and clinical covariates selected and the mean-squared error (*MSE*) using the strategies of ridge, lasso, and elastic net on both the 4 clinicopathological variables and 430 gene expression levels. Moreover, we present results from the lasso and elastic net without penalization on the 4 mandatory covariates, in addition to the ridle. The ridge has the largest *MSE*, suggesting the need for sparse regression. The lasso and elastic net performed similarly with the lasso and elastic net without penalization on mandatory covariates, respectively, in terms of the *MSE*. On the other hand, the ridle performed the best with the smallest *MSE* among all methods. This suggests that the ridle may be advantageous in predicting histologic grades of breast cancer.

Figure 2 depicts the genes selected via sparse regression methods. The ridle encompasses all of the genes selected by the lasso and elastic net with penalization on mandatory covariates, and nearly all of the genes selected by the elastic net (except 1 gene) and lasso without penalization on mandatory covariates (except 2 genes). Two genes are selected only by the ridle: the AREG and TRPM4 genes, which belong to the ErbB signaling pathway and ion-channel family, respectively.

Cells are continuously exposed to stimuli from paracrine and endocrine factors. It is essential that the extracellular signals are interpreted by cell correctly in order to facilitate proper proliferative response. The ErbB family belongs to receptors of the tyrosine kinase family and plays pivotal roles in this process [34]. Members of the ErbB signaling pathway have been suggested as potential therapeutic targets [42]. Initial studies have also suggested that expression levels of AREG (amphiregulin) are associated with larger and more aggressive tumors through cell proliferation [43, 44]. Only the ridle identified AREG as predictive of histologic grades of breast cancers.

The other gene selected only by the ridle is TRPM4 from the ion-channel family. Researches over the past few years have shown that ion channels are involved in the progression and pathology of a myriad of human cancers [39, 45, 46]. In addition, ion channels are known to play critical roles in gene expression, hormone secretion, cell volume regulation, and cell proliferation [47, 48]. The expression levels of ion-channel genes, including TRPM4, have been found to be predictive of and significantly associated with tumor progression [38].

Breast cancer is known to be highly correlated with hormone secretion. Breast tumors that are ER or PgR-positive are much more likely to respond to hormone therapy than tumors that are negative. Many of these may not be related to histologic grades of breast cancer. For example, in a previous study, twenty-four ion-channel genes were found to be differentially expressed between ER-negative and ER-positive tumors [38]. However, in our analysis, we only identified 1 gene, AREG, from the ion-channel family to be predictive of histologic grades of cancer. Thus, many of the 430 breast-cancer related genes may not be predictive of histologic grades but are expected to be highly correlated with the mandatory covariates, i.e. ER and PgR statuses. As suggested by both theoretical and simulation studies, the ridle can be advantageous when mandatory variables are correlated with the irrelevant optional ones. Results from the gene expression analysis further validate and demonstrate the performances of the ridle under this commonly seen scenario.

Table 5 Gene expression analysis on histologic grades of breast cancer

	No. selected $\mathcal{M}$	No. selected $\mathcal{O}$	*MSE*
Ridge	4	430	0.487
Lasso	2	19	0.260
Elastic net	2	14	0.286
$\mathcal{M}$-unpenalized lasso	4	21	0.257
$\mathcal{M}$-unpenalized elastic net	4	7	0.296
Ridle	4	24	**0.239**

The $\mathcal{M}$-unpenalized lasso and $\mathcal{M}$-unpenalized elastic net were performed without penalization on the mandatory covariates. The elastic net and $\mathcal{M}$-unpenalized elastic net are built with alpha=0.2575 and alpha=0.8462, respectively, selected by cross-validation. Numbers of selected mandatory covariates $\mathcal{M}$ and optional variables $\mathcal{O}$, and mean-squared error (*MSE*) are shown. Smallest *MSE* is boldfaced

Discussion

In this article, we proposed the ridle for sparse regression with mandatory covariates. We provided both theoretical and simulation studies that demonstrated the efficacy of our method. In particular, our results suggest that the ridle may outperform the lasso and elastic net when mandatory

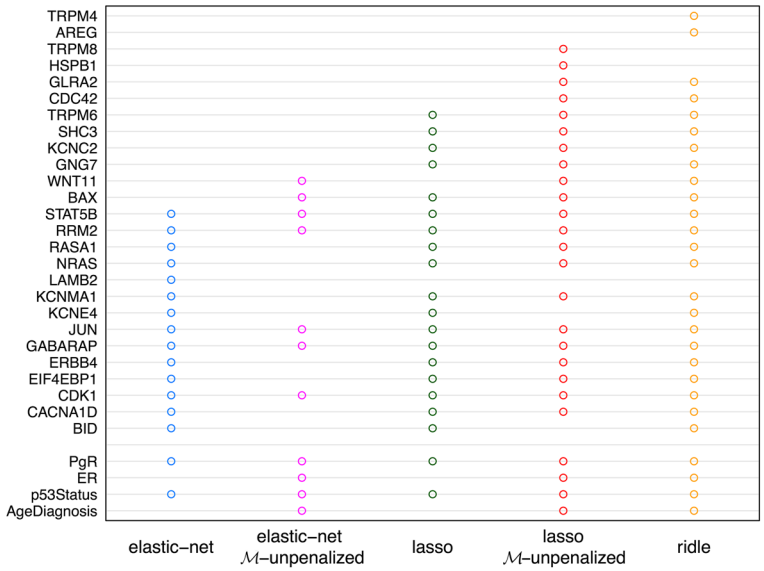

Fig. 2 Selection of genes and clinicopathological variables. PgR, ER, p53Status, and AgeDiagnosis are clinicopathological covariates, whereas all others are genes. The $\mathcal{M}$-unpenalized lasso and $\mathcal{M}$-unpenalized elastic net were performed without penalization on the clinicopathological variables as mandatory covariates

covariates are correlated with the irrelevant optional predictors or are highly correlated among themselves. The ridle can also improve upon performances of the lasso and elastic net when mandatory covariates have small or moderate effects.

We employed the $L1$-norm penalty to induce sparsity on the optional set. This is chosen for its simplicity, computational ease, and successes in a myriad of applications; for example, $L1$-norm penalized regressions have been successfully applied in large-scale genome-wide association [3] and eQTL data studies [49]. However, other sparse regularization methods, such as the SCAD [7], adaptive lasso [8], Dantzig selector [9], etc. can also be utilized in place of the $L1$-norm penalty in (1).

The ridle is related to the elastic net [41] that also employs both the $L1$-norm and $L2$-norm penalties. However, the elastic net applies both penalties upon all coefficients of the optional set, whereas the ridle applies the $L1$-norm to coefficients of the optional set and the $L2$-norm to coefficients of the mandatory set for simultaneous estimation of mandatory covariates while allowing selection for others.

In this article, we applied our method in an interesting application to gene expression analysis where we identified more genes related to tumor grade while incorporating clinicopathological variables as mandatory covariates. In addition, the ridle can be applied in a myriad of other genomic studies where mandatory covariates are routinely required, such as when clinical, demographical, or experimental effects have to be incorporated in regression analysis of genomic data sets.

Conclusions

In this article, we proposed the ridle as a principled sparse regression method for the selection of optional variables while incorporating mandatory ones. Mandatory covariates are routinely encountered in the analysis of genetic-biomedical data. For example, additional covariates describing clinical, demographical or experimental effects need to be included a priori without subjecting them to variable selection. Results suggest that the ridle may outperform current methods when mandatory covariates are correlated with the irrelevant optional predictors or are highly correlated among themselves.

Additional file

Additional file 1: Proof of theoretical results. This file includes the proofs of Theorems 1–3 in "Methods" section. (210 KB PDF)

Abbreviations

NPI: Nottingham Prognostic Index; OLS: ordinary least squares; Ridle: Ridge-lasso hybrid estimator; SCAD: smoothly clipped absolute deviation

Acknowledgements

We would like to thank Gregor Stiglic and Yuan Jiang for their generous comments and improvements made in the article based on their suggestions.

Funding

None.

Availability of data and materials
Dataset used for gene expression analysis on histologic grades of breast cancer in "Results" section can be accessed at the NCBI Gene Expression Omnibus (GEO) repository with GEO accession: gse3294.

Authors' contributions
JZ performed genetic assessment of histologic grades of breast cancer. ZJD planned while JZ performed the simulation studies. ZJD developed the regression method and implemented algorithm. ZJD developed and JZ validated the theoretical results. CHH and ZJD supervised the project. JZ, CHH, and ZJD all participated in project development and writing the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Received: 19 June 2016 Accepted: 6 January 2017
Published online: 25 January 2017

References

1. Liu ZQ, et al. Gene and pathway identification with l(p) penalized bayesian logistic regression. BMC Bioinformatics. 2008;412:1–19.
2. Logsdon BA, Mezey J. Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. PLoS Comput Biol. 2010;6:1001014.
3. Wu TT, et al. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009;25:714–21.
4. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1996;58:267–88.
5. Efron B, et al. Least angle regression. Ann Stat. 2004;32:407–99.
6. Friedman J, et al. Pathwise coordinate optimization. Ann Appl Stat. 2007;1:302–32.
7. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96:1348–60.
8. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101:1418–29.
9. Candes E, Tao T. The dantzig selector: Statistical estimation when p is much larger than n. Ann Stat. 2007;35:2313–51.
10. Trojani M, Contesso G, Coindre J, Rouesse J, Bui N, De Mascarel A, Goussot J, David M, Bonichon F, Lagarde C. Soft-tissue sarcomas of adults; study of pathological prognostic variables and definition of a histopathological grading system. Int J Cancer. 1984;33(1):37–42.
11. Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, Weinberg RA. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet. 2008;40(5):499–507.
12. Gleason DF, Mellinger GT. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J Urol. 1974;111(1):58–64.
13. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Nat Cancer Inst. 2006;98(4):262–72.
14. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12:55–67.
15. Hoerl AE, Kennard RW. Ridge regression: Applications to nonorthogonal problems. Technometrics. 1970;12:69–82.
16. Marquardt DW, Snee RD. Ridge regression in practice. Am Stat. 1975;29: 3–20.
17. Zhao P, Yu B. On model selection consistency of lasso. J Mach Learn Res. 2006;7:2541–67.
18. Stadler N, Buhlmann P, van de Geer S. l1-penalization for mixture regression models. Test. 2010;19:209–56.
19. Tseng P. Coordinate ascent for maximizing nondifferentiable concave functions. Technical Report LIDS-P 1840, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems. 1988.
20. Tseng P. Convergence of block coordinate descent method for nondifferentiable maximation. J Optimiz Theory App. 2001;109:474–94.
21. Friedman J, et al. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
22. Van De Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002;347(25):1999–2009.
23. Bloom H, Richardson W. Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer. 1957;11(3):359.
24. Galea MH, Blamey RW, Elston CE, Ellis IO. The nottingham prognostic index in primary breast cancer. Breast Cancer Res Treat. 1992;22(3):207–19.
25. Engelman JA. Targeting pi3k signalling in cancer: opportunities, challenges and limitations. Nat Rev Cancer. 2009;9(8):550–62.
26. Berns K, Horlings HM, Hennessy BT, Madiredjo M, Hijmans EM, Beelen K, Linn SC, Gonzalez-Angulo AM, Stemke-Hale K, Hauptmann M, et al. A functional genetic approach identifies the pi3k pathway as a major determinant of trastuzumab resistance in breast cancer. Cancer Cell. 2007;12(4):395–402.
27. Levine AJ. p53, the cellular gatekeeper for growth and division. Cell. 1997;88(3):323–31.
28. Sherr CJ, McCormick F. The rb and p53 pathways in cancer. Cancer Cell. 2002;2(2):103–12.
29. Gasco M, Shami S, Crook T. The p53 pathway in breast cancer. Breast Cancer Res. 2002;4(2):70.
30. Skobe M, Hawighorst T, Jackson DG, Prevo R, Janes L, Velasco P, Riccardi L, Alitalo K, Claffey K, Detmar M. Induction of tumor lymphangiogenesis by vegf-c promotes breast cancer metastasis. Nat Med. 2001;7(2):192–8.
31. Jain RK, Duda DG, Clark JW, Loeffler JS. Lessons from phase iii clinical trials on anti-vegf therapy for cancer. Nat Clin Prac Oncol. 2006;3(1):24–40.
32. Kubo M, Nakamura M, Tasaki A, Yamanaka N, Nakashima H, Nomura M, Kuroki S, Katano M. Hedgehog signaling pathway is a new therapeutic target for patients with breast cancer. Cancer Res. 2004;64(17):6071–4.
33. Taipale J, Beachy PA. The hedgehog and wnt signalling pathways in cancer. Nature. 2001;411(6835):349–54.
34. Olayioye MA, Neve RM, Lane HA, Hynes NE. The erbb signaling network: receptor heterodimerization in development and cancer. EMBO J. 2000;19(13):3159–67.
35. Harari D, Yarden Y. Molecular mechanisms underlying erbb2/her2 action in breast cancer. Oncogene. 2000;19(53):6102–14.
36. Downward J. Targeting ras signalling pathways in cancer therapy. Nat Rev Cancer. 2003;3(1):11–22.
37. Clark GJ, Der CJ. Aberrant function of the ras signal transduction pathway in human breast cancer. Breast Cancer Res Treat. 1995;35(1):133–44.
38. Ko JH, Ko EA, Gu W, Lim I, Bang H, Zhou T. Expression profiling of ion channel genes predicts clinical outcome in breast cancer. Mol Cancer. 2013;12(1):1.
39. Kunzelmann K. Ion channels and cancer. J Membr Biol. 2005;205(3): 159–73.
40. Meinshausen N, Bühlmann P. High dimensional graphs and variable selection with the lasso. Ann Stat. 2006;34:1436–62.
41. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B. 2005;67:301–20.
42. Foley J, Nickerson NK, Nam S, Allen KT, Gilmore JL, Nephew KP, Riese DJ. EGFR signaling in breast cancer: bad to the bone. Semin Cell Dev Biol. 2010;21(9):951–60.
43. Ma L, de Roquancourt A, Bertheau P, Chevret S, Millot G, Sastre-Garau X, Espié M, Marty M, Janin A, Calvo F. Expression of amphiregulin and epidermal growth factor receptor in human breast cancer: analysis of autocriny and stromal-epithelial interactions. J Pathol. 2001;194(4): 413–9.
44. Suo Z, Risberg B, Karlsson MG, Villman K, Skovlund E, Nesland JM. The expression of egfr family ligands in breast carcinomas. Int J Surg Pathol. 2002;10(2):91–9.
45. Fiske JL, Fomin VP, Brown ML, Duncan RL, Sikes RA. Voltage-sensitive ion channels and cancer. Cancer Metastasis Rev. 2006;25(3):493–500.
46. Roger S, Potier M, Vandier C, Besson P, Le Guennec JY. Voltage-gated sodium channels: new targets in cancer therapy? Curr Pharm Des. 2006;12(28):3681–95.

47. Camerino DC, Tricarico D, Desaphy JF. Ion channel pharmacology. Neurotherapeutics. 2007;4(2):184–98.
48. Camerino DC, Desaphy JF, Tricarico D, Pierno S, Liantonio A. 4-therapeutic approaches to ion channel diseases. Adv Genet. 2008;64: 81–145.
49. Lee S, et al. Learning a prior on regulatory potential from eQTL data. PLoS Genet. 2009;5:1000358.